View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 114 – 120

Procedia Computer Science

Information Technology and Quantitative Management , ITQM 2013

Supervised Discretization with $GK - \tau$

Wenxue Huang[a], Yuanyi Pan[b,*], Jianhong Wu[c]

[a] *School of Mathematics and Information Sciences, Guangzhou University ,Guangzhou, Guangdong 510006, China*
[b] *InferSystems Corp., 20 Queen Street West, Suite 316, Toronto, Ontario, Canada, M5H 3R3*
[c] *Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada, M3J 1P3*

Abstract

When data are high dimensional and mix-typed while response variable is categorical, an effective executable profile consists of categorical or categorized variables with easily understandable statistics. Many data mining technologies require categorical variables; many have better results by changing continuous variables to categorical variables. Discretizing a continuous variable can be accomplished in either a supervised way or an unsupervised or conventional way. We propose a supervised discretizing method using the Goodman-Kruskal tau (or GK-τ) maximization as the discretization optimization criterion. This optimization is probabilistic averaging effect oriented. An experiment with financial loan application is designed to show the improvement after the discretization. Some technical concerns during the discretization are discussed in this article as well.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: averaging effect; supervised discretization; the GK-tau

1. Introduction

In the world of data mining and machine learning, discretization means categorizing a continuous variable into certain levels. For example, one individual's income can be leveled as low, medium or high; ages can be grouped by five-year steps. Assume that we work with a categorical response variable and explanatory variables among which some or all are continuous variables. For the sake of easily executable profiling, or consistent with descriptive, analytical or averaging-effect oriented proportional prediction, an appropriate discretization is called for. Also, many techniques in this world prefer categorical explanatory variables. The naive Bayes classifying model [21], for instance, is applied in many fields (when the explanatory variables are independent) because of its simple thus quick estimation to the conditional probability. One of its basic assumptions is that the explanatory variable s are all categorical. Another example is the decision tree [22]. Each node in a decision tree is a condition that leads to the next node. The variables involved in each node then have to either be categorical or described as a combination of intervals.

However, many real data sets from industrial applications contain continuous variables such as income, age, interest rate, consumption amount, measure of risk, etc. One of practical solutions to this issue is to treat each distinct value as a member belonging to an appropriate category. An unsupervised, a conventional or non-consistent supervised discretization is not rational in general due to the obvious logical loose or (even) no link.

*Yuanyi Pan Tel.: +1-647-259-9522.
E-mail address: panyuanyi@yahoo.com.

1877-0509 © 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.016

A natural way to group distinct values in a continuous variable is to find out data-driven cutting points that cut the whole range of data into intervals. There are two ways to identify the intervals: with or without a response (or target) variable. Grouping continuous variable with a target (with a given criterion or objective function) is called supervised discretization; while the other (with no link to the response variable) is called unsupervised discretization [5].

Unsupervised discretizations are of interest to data projections with large number of response variables. There are quite a few unsupervised discretazation algorithms. For macro social or economical data, or category product consuming data, an unsupervised discretization algorithm can be based on normal distributions, due to the central limit theorem. Other unsupervised discretization methods include equal interval width and equal frequency intervals [5]. More sophisticated unsupervised methods requires certain quality measures to decide where to cut. One popular measure is the information theoretical entropy-based [4, 6]. The idea is to minimize the entropy in each interval by adjusting the boundaries. Another big family of discretization methods is the clustering technologies [7]. Although most of the methods in this family are applied to multi-dimensional cases, the simplest application of k-mean [19] can also group a one-dimension continuous variable into k parts.

On the other hand, supervised discretization algorithms tune the boundaries by optimizing each interval's coherence [16] associated with a target variable with an optimization criterion. An evaluation function is usually applied to measure the discretization's quality. The typical measures include Chi-square and conditional entropy. The Chi-square based methods include ChiMerge [14], Chi2 [17], Khiops [1] etc. The entropy based methods include the ones in [3], [2], [23], etc. A simpler version of supervised method is Holte's 1R algorithm [9], which rules nothing but a minimum size of each interval with a maximum number of the preferred class.

Which discretization method to be chosen depends on the time computing complexity, the greediness for the accuracy and which framework the method is applied to and the understandability [16] and executability of the result. Nevertheless it is expected that unsupervised discretization methods are faster than supervised methods but less accurate in predicting the target. An experimental evidence by Dougherty et al. [5] shows that entropy-based discretization methods may perform quite well overall regarding the accuracy.

Rather than using entropy-based discretization method, we propose a global-to-global association based measure, Goodman-Kruskal τ, to evaluate the cutting result. Most of times, the entropy-based and the Gini-based are equivalent. The reason we prefer the Gini is not only because the Gini- based GK-τ measure is more directly readable or interpretable than its entropy counterpart [13].

Goodman-Kruskal τ (the GK-τ hereafter) [8] is "a normalized conditional Gini concentration, measuring an averaging effect oriented proportional global-to-global association" [12]. In this article, it is used to choose the optimal cutting point during the greedy searching process.

This paper is organized as follows. Section 2 recalls the definition of GK-τ and introduces its implementation in a discretization framework. Section 3 describes an experiment using a real loan application in banking business. We present some general discussions about discretization, its application and future work in the last section.

2. Discretization with GK-τ

Recall [18, p. 71] that for a categorical explanatory variable X with domain $Dmn(X) = \{1, 2, ..., n_X\}$ and a categorical target variable Y with domain $Dmn(Y) = \{1, 2, ..., n_Y\}$, the association degree of Y on X, denoted by $\tau(Y|X)$ is given by

$$\tau(Y|X) = \frac{\sum_{i=1}^{n_Y}\sum_{j=1}^{n_X} p(Y=i;X=j)^2/p(X=j) - \sum_{i=1}^{n_Y} p(Y=i)^2}{1-\sum_{i=1}^{n_Y} p(Y=i)^2} \tag{1}$$

where $p(\cdot)$ is the probability of an event. In [20] we defined an local-to-local association ${}_i\omega_j$ [10] by

$${}_j\omega_i = p(X=j|Y=i)p(Y=i|X=j) = \frac{(p(Y=i;X=j))^2}{p(X=j)p(Y=i)}$$

where

$$E(p(Y)) = \sum_{i=1}^{n_Y} p(Y=i)^2.$$

We have

$$\tau(Y|X) = \frac{\omega(Y|X) - E(p(Y))}{1 - E(p(Y))} \tag{2}$$

where

$$\omega(Y|X) = \sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} {}_j\omega_i p(Y = Y_i).$$

From the definition above, one can see that ${}_j\omega_i$ measures the interactive predictive power of two scenarios from two variables, that $\omega(Y|X)$ tells us the global(overall) predictive power of an categorical variable X to a dependent variable Y, that $E(p(Y))$ is the overall accuracy rate of the proportional prediction of Y based on its own information, and that $\tau(Y|X)$ is the accuracy lift rate based on information of X over the information of itself.

Please refer to [18] and [12] for more discussions about the previous definitions.

For a given data set with two variables X and Y where X is a continuous variable and Y is a categorical nomial variable with n_y distinct values as defined above. Suppose $C_k = \{c_1, ..., c_k\}$ is a set of distinct real numbers where $c_1 < c_2 < ... < c_k$. Then C_k can used to cut X into maximum $k + 1$ intervals: $(-\infty, c_1], (c_1, c_2], ..., (c_k, +\infty)$. Then τ for a given cutting C_k can be defined as

$$\tau(Y|X(C_k)) = \frac{\sum_{i=1}^{n_Y} \sum_{j=1}^{k+1} p(Y = i; c_{i-1} < X \le c_j)^2 / p(X = j) - \sum_{i=1}^{n_Y} p(Y = i)^2}{1 - \sum_{i=1}^{n_Y} p(Y = i)^2} \tag{3}$$

where $c_0 = -\infty$ ad $c_{k+1} = +\infty$. Thus $\tau(Y|X(C_k))$ measures the overall predictive power of a cutting C_k.

We propose a greedy searching scheme for cutting points in a continuous categorical variable X with respect to a categorical variable Y as follows.

1. Create the initial cutting points C_K to X by an unsupervised discretize method;
2. Set the initial number of chosen cutting points, m, named 0;
3. Loop the following steps until the condition is met;
   (a) Counting the chosen cutting points in C_K that work as the boundaries, say m;
   (b) If $m \ge \theta_b$ where θ_b is the predefined maximum number of intervals, stop the loop;
   (c) Otherwise, suppose $B_m = \{b_1, ..., b_m\}$ contains the chosen boundaries, choose the next boundary b_{m+1} such that

$$b_{m+1} = \arg \max_{b \in C_K \setminus B_m} \tau(Y|(X(B_m \cup \{b\})))$$

Please note that τ is equivalent to $\omega(Y|X)$ in the previous steps since $E(p(Y))$ is always the same during the whole process.

Basically, this scheme checks all the available cutting points, finds out the one plus which the chosen cutting points generate the biggest τ and stops only when the maximum number of intervals is reached. It is apparently not a fancy approach but it has been widely used in various concretization algorithms according to [16]. Besides, the major purpose of this article is to illustrate and verify the application of τ in discretization after all.

3. Experiment

The data set in this experiment is a real loan application data set discussed in [20] and [12]. It has 650 rows with both continuous variables and categorical variables. We choose *On-Time* (repaying of loan) as the target variable and *Income* as the continuous response variable. *On-Time* is a binary variable with values of 0 and 1. When 0 indicates the customer who didn't repay the loan on time and 1 means the contrary, it is mostly 0 that is the targeted class with smaller proportion as 0.1. The explanatory variable was categorized as *low*, *average* and *high* in both literatures by industrial protocal as this: when the income is less than or equal to 30000 a year, it is low; when it is more than 80000, it is high; otherwise, it is average. Apparently it is discretized in an unsupervised way. Despite this discretization method's wide and effective application in many fields, Huang et. al. found in [12] that it has very low association with this specific target variable in this specific data set. The goal of this

experiment is to prove that τ defined in the previous section can be used to discretize the variable *Income* to increase the association and better predict *On-Time*.

To achieve this goal, the data set is randomly splitted into two parts: the one part is used to discretize *Income* and to be trained for predicting the given response variable *On-Time*; *On-Time* in the second part is then predicted; the real and the predicted values of *On-Time* in the second part are compared to evaluate the prediction performance. The first part is usually called the learning set and the second part is called the test set.

Although lift curve based indices measure the performance for most rare event targeting [11], we will use confusion matrix based criterion to evaluate the result for binary response variable for a general purpose. In cases that the response variable is multinomial, we recommend the recently introduced association matrix by Huang et al ([12]). Even each value in this variable is assigned a different score, the scorings are hardly supportive enough to correctly estimate the indices introduced in [11] including G, G_{ph} and G_{ip}.

The purity with respect to the target values in each interval is higher on average if the final number of intervals is bigger. It means that the overall predictive power is in general bigger. Thus the final number of intervals after the supervised discretization is chosen as the same 3 as the default unsupervised discretization to fairly compare the results.

Another issue of the discretization in Section 2 is the initial intervals. Since it is not the major concern in this article, we deliberately choose 20 boundaries, equally distributed from 10000 to 110000.

Since the total number of rows in the whole data set is so low, the sampling and splitting result may significantly influence the result. So the experiment is repeated 100 times to average out the sampling variance.

In summary, the experiment includes the following steps.

1. The data set is split into two parts;
2. *Income* in the learning set is discretized by the method introduced in Section 2 and by the default industrial protocal;
3. Calculate the conditional probabilities of supervised and unsupervised categorical values in *On-Time*;
4. Predict *On-Time* in the test set by the conditional probabilities from Step 3;
5. Evaluate the prediction results by confusion matrix based statistics;
6. Repeat Step 1 to Step 5 for 100 times.
7. Average out the statistics in 5.

Two simple ways of predicting binary variables are tested in this article. The first one is to predict the target value with bigger conditional probability. It is applied mostly to the case that the two values in the binary variable have approximately the same proportion. The second one is to predict the target value as the rare one when the conditional probability is greater than average. It is used mostly in the rare events targeting where the first predicting most likely chooses only the dominant class, which invalidates all the preprocessing concerns including discretization.

Given that x_j is a category in the categorical variable and $f(x_j)$ is the prediction function, the previous prediction approaches, noted as f_1 and f_2 respectively, are described as follows.

$$f_1(x_j) = \begin{cases} 1, & \text{if } p(Y=1|x_j) > p(Y=0|x_j); \\ 0, & \text{otherwise .} \end{cases}$$

$$f_2(x_j) = \begin{cases} 1, & \text{if } p(Y=1|x_j) > p(Y=1); \\ 0, & \text{otherwise .} \end{cases}$$

The experiment runs under SAS. The program, the raw and processed data sets are available upon request to the authors. We present the test result as follows.

The first table contains the basic statistics for the 100 times of sampling and splitting. The total number or rows, the minimum/maximum/average/standard deviation of *Income* in each sampling are collected into *n*,*min*,*max*,*mean* and *std*. The averages and standard deviations of these 100 groups of statistics are then summarized into row *mean* and *std* in the table. One can see that that the learnings sets and their corresponding test sets have almost the same distribution, which ensures that the test doesn't deviate from the learning too much.

Table 1. Sampling statistics to *Income* : learning set and test set

Learning	n	min	max	mean	std	median
mean.	325	8,555	208,588	61,289	29,420	57,238
std	14	2,077	15,901	1,224	1,376	1,472
Test	n	min	max	mean	std	median
mean	325	8,049	205,812	61,170	29,284	57,064
std	14	1,632	19,936	1,210	1,391	1,413

Table 2. Discretization results: unsupervised v.s. supservised

	Supervised discretization			Default discretization		
	Boundary 1	Boundary 2	$\omega(Y\|X)$	Boundary 1	Boundary 2	$\omega(Y\|X)$
min	15,000	40,000	0.7809	30,000	80,000	0.7798
max	85,000	110,000	0.8644	30,000	80,000	0.8609
mean	34,450	72,350	0.8248	30,000	80,000	0.8227
std	14,388	20,676	0.0188	0	0	0.0186
median	30,000	75,000	0.8249	30,000	80,000	0.8232

The second table shows the statistics for the boundaries found in the supervised and the default discretizations. The 100 pairs of boundaries and 100 $\omega(Y|X)$s are collected and the statistics to them are calculated in the table above. One can see the supervised one generates better association.

Table 3. Test results: unsupervised v.s. supervised

Supervised discretization to predict method f_1											
	FP	TP	TN	FN	recall	precision	F1	accuracy	p_{neg}	r_{neg}	$F1_{neg}$
mean	32.42	292.68	0.05	0	0.9998	0.9003	0.9474	0.9001	0	0	null
std	4.137	13.0096	0	0.5	0.0017	0.012	0.0066	0.0119	0	0	null
Default discretization to predict method f_1											
	FP	TP	TN	FN	recall	precision	F1	accuracy	p_{neg}	r_{neg}	$F1_{neg}$
mean	32.42	292.73	0	0	1.	0.9003	0.9475	0.9003	0	0	null
std	4.137	13.0166	0	0	0.	0.012	0.0066	0.012	0	0	null
Supervised discretization to predict method f_2											
	FP	TP	TN	FN	recall	precision	F1	accuracy	p_{neg}	r_{neg}	$F1_{neg}$
mean	19	14	145	147	0.5031	0.9231	0.6103	0.5105	0.1223	0.583	0.1905
std	9	9	74	75	0.2522	0.0231	0.2192	0.2017	0.0292	0.2573	0.0301
Default discretization to predict method f_2											
	FP	TP	TN	FN	recall	precision	F1	accuracy	p_{neg}	r_{neg}	$F1_{neg}$
mean	28	5	209	84	0.2861	0.9499	0.4361	0.3429	0.1172	0.8551	0.2056
std	4	3	18	18	0.0587	0.02	0.068	0.047	0.0149	0.076	0.0236

The third table shows how supervised discretization improves the prediction. As collected in the previous tables, all 100 groups of statistics are calculated; the averages and standard deviations to each statistics are summarized in row *mean* and row *std*. Please note the statitics in this table are defined below.

FP = ♯ of rows predicted as 1 while the real value is 0;

$TP = \sharp$ of rows predicted as 1 while the real value is 1;

$TN = \sharp$ of rows predicted as 0 while the real value is 0;

$FN = \sharp$ of rows predicted as 0 while the real value is 1;

$$recall = \frac{TP}{FN + TP};$$

$$precision = \frac{TP}{FP + TP};$$

$$F1 = 2\frac{precision \times recall}{precision + recall};$$

$$accuracy = \frac{TP + TN}{FP + TP + TN + FN};$$

$$p_{neg} = \frac{TN}{TN + FN};$$

$$r_{neg} = \frac{TN}{TN + FP};$$

$$F1_{neg} = 2\frac{p_{neg} \times r_{neg}}{p_{neg} + r_{neg}};$$

One can refer to [15] for detailed discussion about the first 6 statistics. The last 3 are introduced to address the special need in this experiment, in which negative class 0 is the class with less proportion but probably of more concerns. p_{neg},r_{neg}, $F1_{neg}$ are the precision, the recall and $F1$ regarding 0 respectively;

According to Table 3, the predicting performance by f_1 after the supervised discretization is statistically the same as that after the unsupervised discretization. On the other hand, f_2 shows that the supervised discretization performances better overall than the unsupervised one. Although the superiority focus most on the dominant class and the precision of 0 is also better under the supervised discretization, the recall and $F1$ of the rare class 0 is better under the default unsupervised discretization approaches.

4. Discussion and future work

We present a new supervised discretization algorithm using the global-to-global association measure, the GK-τ. We believe that it is more interpretable than the entropy based one. It also shows better predicting result than the unsupervised discretization method by an experiment to a real industrial data of loan application in banking. However, it is no surprise to find out that the optimal (or conditional mode based) prediction could ruin everything even the pre-processing steps are perfect. In our case, it means that unsupervised or supervised, it makes no differences when the target binary variable is imbalanced and the conditional mode based optimal prediction does not concerns this imbalance at all. In this case, the proportional prediction realized via Monte-Carlo simulation can faithfully reflect the conditional distributions and the lifts. For multinomial targets and for the conditional mode based optimal predictions oriented supervised discretizations should be an interesting topic to be studied.

The experiment in this article only uses one categorical variable with only three categories. It limits our choice of predicting methods. One future work is to test the proposed method with more various scenarios. For example, it would be very interesting to find out how the proposed method works under feature selection in a high-dimentional data set; how the proposed method works for a balanced binary target variable or an imbalanced binary target variable with 1 as the rare class but with more concerns; how the proposed method performances under other predicting method,e.g, lift curve based prediction; etc.

References

[1] Boulle, M., 2004. Khiops: A statistical discretization method of continuous attributes. Machine Learning 55 (1), 53–69.
[2] Catlett, J., 1991. On changing continuous attributes into ordered discrete attributes. In: Machine LearningEWSL-91. Springer, pp. 164–178.
[3] Chiu, D., Cheung, B., Wong, A., 1990. Information synthesis based on hierarchical maximum entropy discretization. Journal of Experimental & Theoretical Artificial Intelligence 2 (2), 117–129.
[4] Chmielewski, M. R., Grzymala-Busse, J. W., 1996. Global discretization of continuous attributes as preprocessing for machine learning. International journal of approximate reasoning 15 (4), 319–331.
[5] Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. Morgan Kaufmann Publishers, Inc., pp. 194–202.
[6] Fayyad, U., Irani, K., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the International Joint Conference on Uncertainty in AI.
[7] Gan, G., Ma, C., Wu, J., 2007. Data clustering: Theory, algorithms, and applications (asa-siam series on statistics and applied probability). 2007. Society for Industrial & Applied Mathematics, USA.
[8] Goodman, L., Kruskal, W., 1954. Measures of association for cross classifications*. journal of the American Statistical Association 49 (268), 732–764.
[9] Holte, R., 1993. Very simple classification rules perform well on most commonly used datasets. Machine learning 11 (1), 63–90.
[10] Huang, W., Pan, Y., Wu, J., 2012. Goodman–kruskal measure associated clustering for categorical data. International Journal of Data Mining, Modelling and Management 4, 334–360.
[11] Huang, W., Pan, Y., Wu, J., 2012. Performance measures of rare events targeting. International Journal of Data Analysis Techniques and StrategiesTo appear.
[12] Huang, W., Shi, Y., Wang, X., 2011. Nominal association vector and matrix. arXiv preprint arXiv:1109.2553.
[13] Huang, W., Vainder, M., 2004. Dependence degree and feature selection for categorical data. Workshop on Data Mining Methodology and Applications at The Fields Institute.
[14] Kerber, R., 1992. Chimerge: Discretization of numeric attributes. In: Proceedings of the tenth national conference on Artificial intelligence. AAAI Press, pp. 123–128.
[15] Kohavi, R., Rovost, F., 1998. Glossary of Terms. Machine Learning 30, 271–274.
[16] Kotsiantis, S., Kanellopoulos, D., 2006. Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering 32 (1), 47–58.
[17] Liu, H., Setiono, R., 1995. Chi2: Feature selection and discretization of numeric attributes. In: Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on. IEEE, pp. 388–391.
[18] Lloyd, C. J., 1999. Statistical analysis of categorical data. A Wiley-Interscience publication. Wiley, New York, NY, USA.
[19] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. California, USA, pp. 281–297.
[20] Olson, D., Shi, Y., 2007. Introduction to business data mining. McGraw-Hill/Irwin.
[21] Rish, I., 2001. An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. pp. 41–46.
[22] Safavian, S., Landgrebe, D., 1991. A survey of decision tree classifier methodology. Systems, Man and Cybernetics, IEEE Transactions on 21 (3), 660–674.
[23] Ting, K., 1994. Discretization of continuous-valued attributes and instance-based learning. Basser Department of Computer Science, University of Sydney.